

Community core detection in transportation networks

Vincenzo De Leo,¹ Giovanni Santoboni,² Federica Cerina,^{1,3,4}

Mario Mureddu,^{1,3,4} Luca Secchi,¹ and Alessandro Chessa^{1,4,5}

¹*Linkalab, Complex Systems Computational Laboratory, Cagliari, Italy*

²*AECOM P3 Advisory, 2201 Wilson Boulevard, Suite 800, Arlington VA, 22201, USA*

³*Department of Physics, University of Cagliari, Italy*

⁴*IMT Institute for Advanced Studies Lucca, Piazza S. Ponziano 6, 55100 Lucca, Italy*

⁵*Institute for Complex Systems, CNR UOS Department of Physics, University of Rome "La Sapienza", Rome, Italy*

This work analyzes methods for the identification and the stability under perturbation of a territorial community structure with specific reference to transportation networks. We considered networks of commuters for a city and an insular region. In both cases, we have studied the distribution of commuters' trips (*i.e.*, home-to-work trips and viceversa). The identification and stability of the communities' cores are linked to the land-use distribution within the zone system, and therefore their proper definition may be useful to transport planners.

PACS numbers: 89.75.-k, 87.23.Ge, 05.70.Ln, 89.75.Hc, 89.20.Hh, 05.10.-a

I. INTRODUCTION

Many Complex Systems can be modeled as networks, in which vertices are the entities of interest in the system under investigation and edges are the relations between couple of vertices/entities. For example in the World Wide Web the vertices are the web pages and the edges are the hyperlinks (in this case the network is directed and we have arcs instead of simple edges). Intuitively, not all vertices and edges have equal roles within a large-scale network; some vertices may be of some importance for the distribution of traffic in the network, and the edges that carry most of the traffic do so because they connect "groups" of vertices that are particularly important within the network. The scope of this paper is to understand the nature of these "groups", their "community structure" or "clustering", and find ways to determine the importance of vertices inside each community, revealing its inner hierarchy. The community structure of a network is a topic that has been comprehensively treated in [1].

The first problem of graph clustering is one of definition. Although the concept is intuitive, it is not defined in a rigorous way, as there is no definition of community boundary, or a unique way of determining whether a particular edge is part of a community and not of another. Therefore, as pointed out in [1], communities are algorithmically defined, *i.e.*, they are the final product of the algorithm, without a precise *a priori* definition.

This paper analyses methods for the identification and the stability of a community structure using two networks from the field of transportation. The first network is a regionwide network of commuting trips in the insular region of Sardinia, in Italy. The second network is a network of daily commuting trips in the metropolitan area of Atlanta, USA. In both cases, we have studied the distribution of commuting trips, *i.e.*, home-to-work trips and viceversa. The choice was determined by the fact that trips of these types are clearly defined to planners, be-

cause their correlation to the land-use is well understood, necessarily tied to the population of the origin zone and the employment of the destination zone.

The field of transportation is a natural choice for the definition of a community structure, though the field itself has some inherent limitations. On a practical matter, the measurement of important traffic variables is lengthy and expensive. For once, different methods to count traffic volumes return different answers, especially in the identification of commercial vehicles [2]. Additionally, the development of a regionwide origin-destination (OD) matrix at the zone level is a long and costly procedure; in particular the matrix of the metropolitan area used in this study has been derived after a year-long survey process, and the final OD matrix is assembled by weighting a matrix of survey responses according to the population of the areas where the participants live. A second calibration stage is generally done to test whether the OD matrix obtained assigns traffic compatibly with the traffic on the major highways of the study area; as a result of this process, the trip distribution and assignment may work well *globally*, but larger discrepancies may persist *locally*. Finally, during the time occurred to carry out this process, conditions on the ground may have already changed, since the land-use of an area is constantly changing, therefore creating discrepancies in the final OD matrix.

Notwithstanding these inherent difficulties, the identification of communities within a metropolitan area network still holds great importance. First, the formation of communities in a network is a byproduct of land-use development. Land-use development occurs for a number of reasons (service maximization, profit, etc), and the location for development is chosen according to the optimization in terms of different variables, like price of land, proximity to transit, regulation, that are however variables related to each zone/vertex of the system. For example, demand for transport between two vertices may lead to the opening of a new edge (*e.g.*, a new bus route, a new road), which in turn may lead to more demand

for transport (in the form of "induced demand", [3, 4]). The community structure is not solely a function of the attributes of each zone/vertex, but also of the network arrangement, hence it forms a more comprehensive measure of the importance of a group of zones as a subsection of the zone system.

It is important to know which vertices are the most relevant from the point of view of the internal stability of a community and the overall partition structure. We will see in the next section that this idea is at the cornerstone of the community stability. In other fields the problem has been studied in terms of network breakdown, which has found applications in the accessibility of a transportation network for flood damage. Knowledge of community structure can serve planners in the situation of natural disasters to predict the onset of network breakdown, as studied in [5]. In other fields, it has been applied to the identification of crucial edges in a web network under cybernetic attack [6–8].

II. MATERIALS AND METHODS

A. Community detection and modularity

There are now many community detection methods [1] and the most popular is the modularity optimization introduced by Newman and Girvan [9]. This method has various drawbacks, the most important of which is the existence of a resolution limit [10] which prevent it to detect smaller modules, but has also the advantage of being easy to implement. The modularity function that needs to be optimized is defined as [11]:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(C_i, C_j) \quad (1)$$

where the sum is over all the node pairs, A is the adjacency matrix, m is the total number of edges and P_{ij} is the expected number of edges between the vertices i and j for a given null model. The function will result in a null contribution for couples of vertices not belonging to the same community ($C_i \neq C_j$). For an unweighted network, the choice $P_{ij} = k_i k_j / 2m$ equates to taking as a null model a random network with the same degree sequence as the original network.

To optimize the modularity we used the Louvain algorithm [12] based on two steps that are repeated iteratively until a global maximum is reached. In the first step we create a network partition where the number of communities is equal to the nodes number. Then, the algorithm iterates over all nodes and computes for each node the modularity gain within the communities of its neighbors; a node movement is maintained if it leads to a positive variation in modularity. The iteration is repeated until a local maximum is reached, that is until there is not any other move that lead to an increase in modularity.

In the second step the algorithm creates a new network

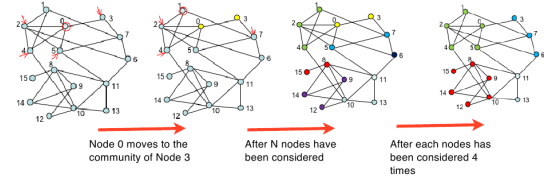


FIG. 1: (Color online) This figure shows an example of the first step execution over a network with 15 nodes: at the beginning all nodes are isolated (left), then the algorithm start to merge several nodes together (center) until the local maximum is reached (right) (after Blondel et al. [12])

whose nodes are the communities; the total weight of the links between communities is the total weight of the links between the nodes of these communities. Typically the nodes number diminishes drastically at this step and this ensures the rapid convergence of the algorithm for large networks.

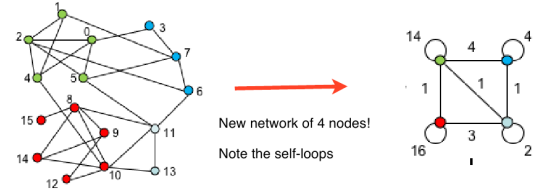


FIG. 2: (Color online) This figure shows an example of the second step where it is possible to note the creation of self links associated to the communities internal connections(after Blondel et al. [12]).

The main problems of all algorithms for community detection is the fact that the community definition does not provide any information about the importance of a node inside its own community. Nodes of a community do not have all the same importance for the community stability: the removal of a node in the "core" of a network affects the partition much more than the deletion of a node that stays on the edge of the community (*i.e.* a node connected in the same way with nodes internal and external to its community). The purpose of the following section is to develop a novel way for detecting cores inside communities by using the properties the of modularity function.

B. dQ analysis for cores detection in a partition

By definition, if the modularity associated to a network has been optimized, every perturbation in the partition leads to a negative variation in the modularity (dQ). If we move a node from a partition we have $M - 1$ possible choices (with M the number of communities) as possible targets for the new host community of this node. We decided to define the dQ associated to each node as the smallest variation in absolute value (or the closest to 0

since dQ is always a negative number) for all the possible choices and this is in our view a measure of how that node is internal in its community.

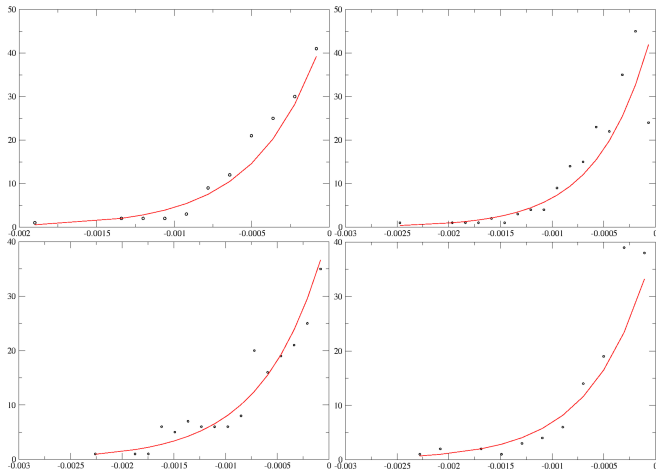


FIG. 3: (Color online) dQ frequency plots relative to 4 communities detected for the city of Atlanta, GA. The correlation coefficients of the exponential fits are (from top right to bottom left, respectively) 0.956, 0.946, 0.937 and 0.933. In general, these distributions are the typical dQ frequency distribution inside a community (provided there are enough nodes to perform an exponential fit).

Fig. 3 shows the typical dQ frequency distribution of nodes inside a community; the data points were fitted using a decaying exponential form $\exp(-x/\ell)$ with typical length ℓ . The typical length ℓ defines a starting point to discriminate the core nodes. For practical purposes, the threshold value $d_{thr} = 2\ell$, is an appropriate boundary value to differentiate between core nodes (the ones below the threshold) and the border nodes (the peripheral nodes).

Fig. 4 shows the cores detected for the city of Atlanta, GA, using the method described above.

III. DATASETS

A. Sardinian Inter-municipal Commuting Network

Sardinia is the second largest Mediterranean island with an area of approximately 24,000 square kilometers and 1,600,000 inhabitants. At the date of 1991, the island was partitioned in 375 municipalities, the second simplest body in the Italian public administration, each one of those generally corresponding to a major urban centre (in Figure 5 we report the geographical distribution of the municipalities). For the whole set of municipalities the Italian National Institute of Statistics [13] has issued the origin-destination table (OD) corresponding to the commuting traffic at the inter-city level. The OD is constructed on the output of a survey about commuting behaviors of Sardinian citizens. This survey refers

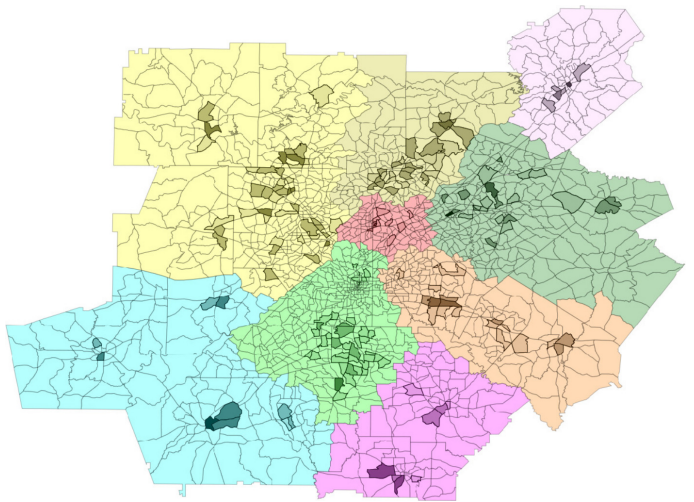


FIG. 4: (Color online) Cores detected for the city of Atlanta, GA, using a threshold equal to double the typical length of the exponential distribution of the dQ frequencies.

to the daily movement from the habitual residence (the origin) to the most frequent place of employment (the destination): the data comprise both the transportation means used and the time usually spent for displacement. Hence, OD data give access to the flows of people regularly commuting among the Sardinian municipalities. In particular we have considered the external flows $i \rightarrow j$ which measure the movements from any municipality i to the municipality j and we will focus on the flows of individuals (workers and students) commuting throughout the set of Sardinian municipalities by all means of transportation. This data source allows the construction of the Sardinian inter-municipal commuting network (SMCN) in which each node corresponds to a given municipality and the links represent the presence of a non-zero flow of commuters among the corresponding municipalities.

The standard mathematical representation of the resulting network is provided by the adjacency matrix A of elements (a_{ij}) . The elements on the principal diagonal (a_{ii}) are set equal to zero, since intra-municipal commuting movements are not considered here. Off-diagonal terms a_{ij} are equal to 1 in the presence of any non-zero flow between i and j ($i \rightarrow j$ or $j \rightarrow i$) and are equal to 0 otherwise. The adjacency matrix is then symmetric and describes regular bi-directional displacements among the municipalities. The adjacency matrix contains all the topological information about the network but the dataset also provides the number of commuters attached to each link. It is therefore possible to go beyond the mere topological representation and to construct a weighted graph where the nodes still represent the municipal centres but where the links are valued according to the actual number of commuters. Analogously to the adjacency matrix A , we thus construct the symmetric weighted adjacency matrix W in which the elements w_{ij} are computed as the sum of the $i \rightarrow j$ and $j \rightarrow i$ flows between

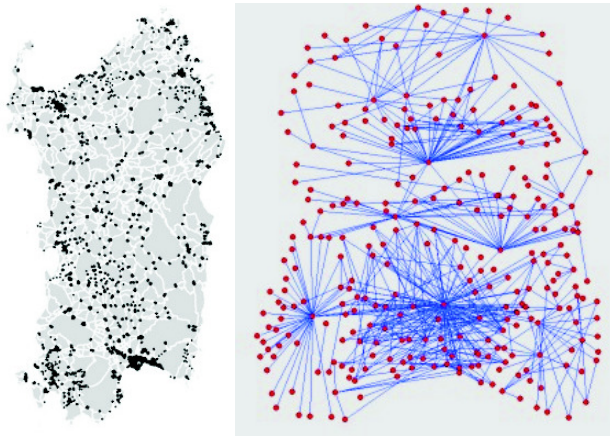


FIG. 5: (Color online) Geographical versus topologic representation of the the Sardinian inter-municipal commuting network (SMCN): the nodes (red points) correspond to the towns, while the links to a flow value larger than 50 commuters between two towns.

the corresponding municipalities (per day). The elements w_{ij} are null in the case of municipalities i and j which do not exchange commuting traffic and by definition the diagonal elements are set to zero. According to the assumption of regular bi-directional movements along the links, the weight matrix is symmetric and the network is described as an undirected weighted graph. The weighted graph provides a richer description since it considers the topology along with the quantitative information on the dynamics occurring in the whole network.

B. ARC Network

The Atlanta Regional Commission (ARC) maintains a network model for land use purposes of the metropolitan area of the city of Atlanta, in the State of Georgia, USA. The ARC travel demand model is designed to represent the state of the practice in travel demand modeling and to meet all modeling requirements in the US EPA Transportation Conformity Rule. Further details on the arrangement of zones are reported in [14].

The main data source for the calibration of the travel demand models was a household travel survey of eight thousand households conducted for the ARC from April 2001 through April 2002. The household survey data was the main source of data for developing the trip generation and distribution model. The trip generation model is a fairly unique trip based model in that it estimated the frequency a person will make trips, by the purpose of the trip, and then applies this frequency to individual persons to determine the total amount of travel made by the residents of the region. Therefore, as in the case of the SMNC network, the trips reported in the ARC model are produced by a trip generation model, which is calibrated according to the result of a survey. The calibration is

achieved by matching the trip length, frequency and by evaluating geographic area biases (*e.g.*, natural features, political or service delivery boundaries, etc).

The work presented in this paper is centered on the activity of commuters, which in the ARC model are described as Home Based Work (HBW) trips. It is commonplace to describe such trips as trips made for the purpose of work and which either begin or end at the traveler's home. This is a typical trip purpose that is related to the employment at the destination zone and population/household income of the traveler or the household at the origin zone. Mode details on the nature and calibration of the HBW demand and distribution model can be found in [14] for this specific model. The nature of the relationship between demand for travel and land-use are further explored in the modeling review works by Wilson [15] and Batty [16].

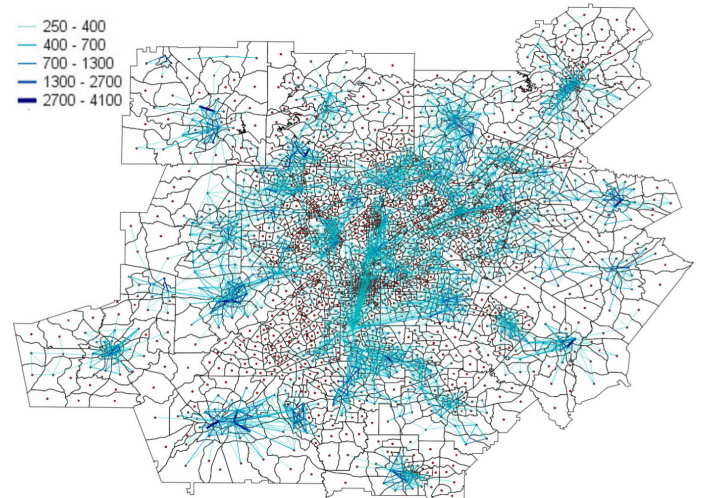


FIG. 6: (Color online) Extension of the zone system in the ARC model. Only the links with a weight greater than 250 have been shown. Each point is a centroid of a TAZ.

A number of socioeconomic variables are recorded in the ARC model, which are of importance for planning purpose and as inputs to the trip generation and demand growth algorithms. The figures below show, in order, the gradient plots of population and employment per zone, as recorded in the nationwide Census 2010. Darker zones indicate higher value for the corresponding variable.

Figure 7 shows the gradient plot of the zone population. Population is seen in this figure as being scattered around the center that forms the core of the downtown area.

Figure 8 shows the gradient plot for the zone employment, measured as the number of jobs located in the zone the variable refers to. Employment is seen in this figure as primarily located in the downtown zones (which are quite small in size) plus other job centers in the suburban metropolitan areas.

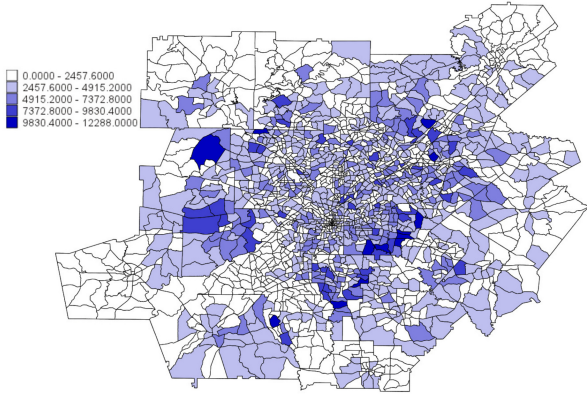


FIG. 7: (Color online) Gradient plot for Population in the ARC model.

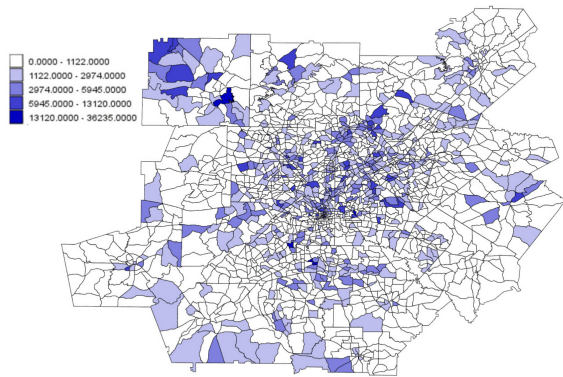


FIG. 8: (Color online) Gradient plot for Employment in the ARC model.

IV. RESULTS

The sequence of charts that follow describes the correlation of the quantity dQ and the various socioeconomic variables that are available for analysis.

The table below shows the result of correlation analysis between the computed dQ and the in-strength of the various zones in the SMCN network. For the sake of clarity, the Sardinian and ARC networks are in principle directed, as previously described in III, and the in-strength has been computed starting from these original networks. On the contrary, the community detection has been performed using undirected networks obtained from the directed ones by summing up the weights of incoming and outgoing links. The correlation results shown in the table I only give a overall picture of the quality of correlation between traffic and community structure. Figures 9-10 show the geographic distribution of the gradients of dQ values across the zone system. Figure 9 shows the values of dQ arranged by color (darker color indicates higher value). Higher dQ indicates that the zone under

in-strength	Correlation
Employment	0.984
Academic	0.977
Both	0.984

TABLE I: Results of correlation analysis between dQ and the in-strength related to particular segments of the traveling population in the SMCN network.

investigation is more to the center of a community than the zones with lighter color. The data in Figure 9 shows that the two likeliest centers of a community (the two darkest zones in the figure) are not both centers of population and/or employment, nor are all large centers of population and/or employment necessarily key zones to the definition (and for its definition, stability) of a community. In other words, community and socioeconomic activity are not on a one-to-one relationship, and it is not always possible to imply a ranking of one of these quantities with respect to the other and viceversa.

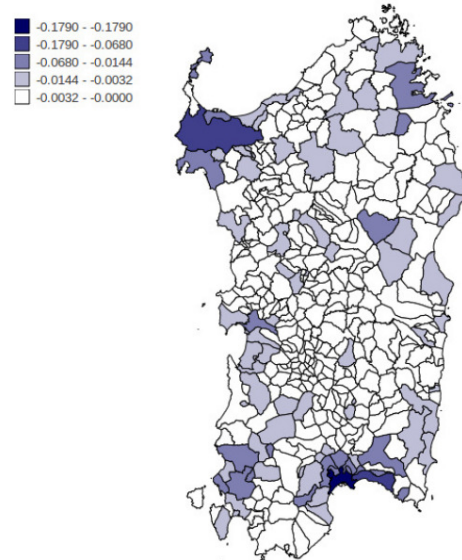


FIG. 9: (Color online) dQ plot for the network related to Employment in the SMCN network.

Figure 10 (right) below shows what the communities identified look like with respect to the political subdivisions of the island of Sardinia, the provinces that corresponds to the NUT3 regions in the international classifications (left). To put this result in context, it is important to note that the present political subdivision in eight provinces took effect in 2005 after a law passed in 2001 raised the number of provinces from the original number of four. Therefore, at the time the ISTAT data was collected (2001), Sardinia was subdivided politically in four provinces, hence the results of the modularity analysis showed that at least seven communities existed, subdivided geographically roughly along the lines of the

boundary of the new (and present time) provinces. The two subdivisions, "topological" the first, political the second, are remarkably alike, suggesting that either the political subdivision was designed to accomodate the arrangement of commuting movements, or the topological subdivision is a result of ease of movement within a (not yet established) political subdivision.

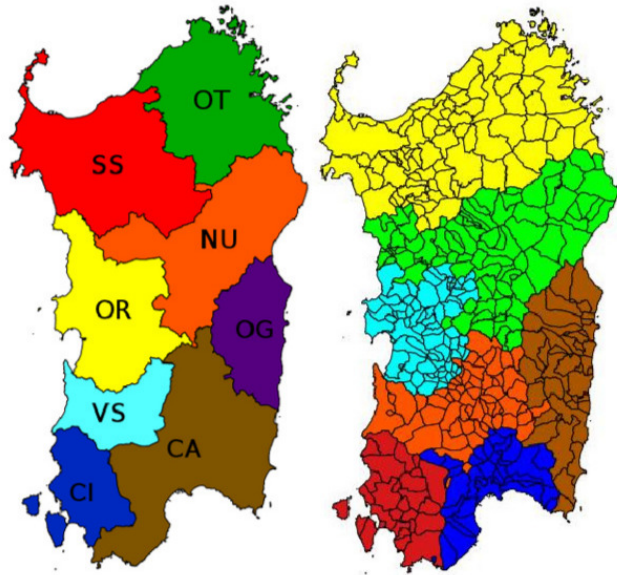


FIG. 10: (Color online) A comparison between the current provincial division (CA = Cagliari, CI = Carbonia-Iglesias, VS = Medio Campidano, OR = Oristano, OG = Ogliastra, NU = Nuoro, SS = Sassari and OT = Olbia-Tempio) of the Sardinia region, Italy, and the result of the community detection.

Finally, it is worth noting that, according to the results of a regional referendum in May 2012, the four new provinces established in according to the 2001 law will be abolished starting March 2013.

Table II shows the result of the correlation between in-strength, dQ and employment for the ARC network. Correlation with employment is quite poor while, as in the case of the SMCN network, correlation with the in-strength is quite good. It is instructive then to see the geographic arrangement of the communities and other features of the network. Figure 11 shows the dQ dis-

Variable	Correlation
in-strength	0.782
Employment	0.052

TABLE II: Results of correlation analysis between dQ and various variables in the ARC network.

tribution for the ARC network. Darker zones indicate zones with higher dQ , and the darkest zones can be considered as the center of a community. Figure 12 show

(color-coded) the community boundaries. The correla-

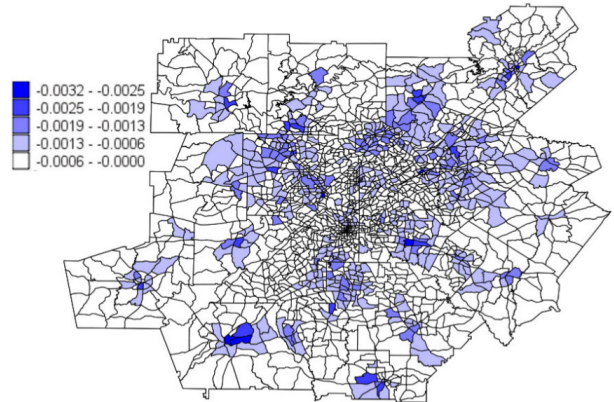


FIG. 11: (Color online) dQ plot for the ARC network.

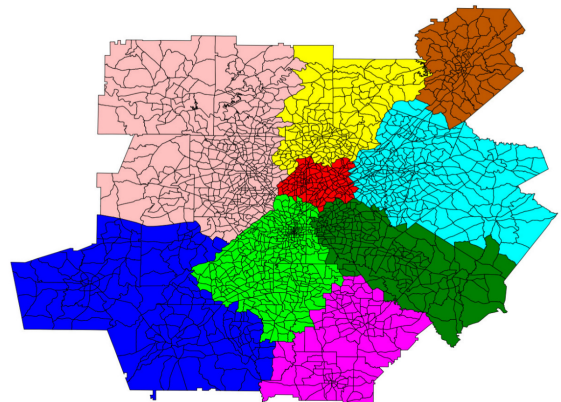


FIG. 12: (Color online) dQ and community boundary plot for the ARC network

tion between dQ and in-strength is explored by means of the Figure 13, which shows a correlation of almost 0.8. As per the case of the SMCN network, community and socioeconomic activity are not on a one-to-one relationship, and it is not always possible to imply a ranking of one of these quantities with respect to the other and viceversa.

V. DISCUSSION

The two case studies that have been the subject of this analysis showed that community structure coming from the networks analysis with its cores definitions, and socioeconomic activity are not on a one-to-one relationship, and it is not always possible to imply a ranking of one of these quantities with respect to the other and viceversa. Hence, the "community" is a distinct mathematical object with its own land-use meaning that contains

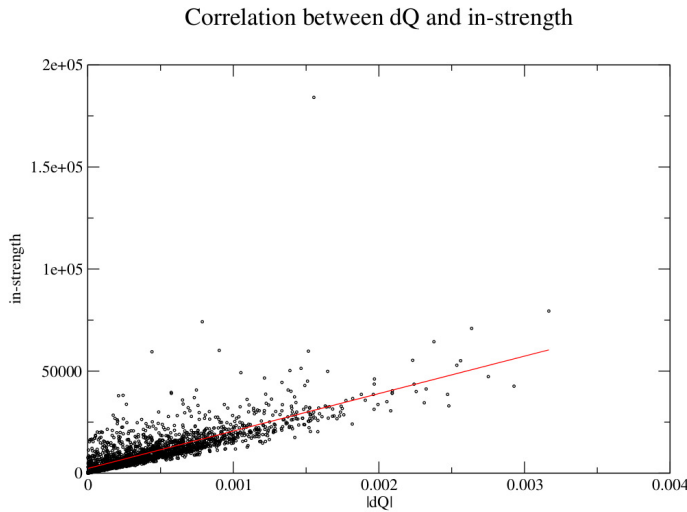


FIG. 13: (Color online) The correlation between dQ and in-strength is equal to 0.78.

some valuable information to be exploited. Correlation

between the community stability (expressed in dQ value) and socioeconomic variables only tells part of story, while the remaining contribution to the community stability is to be found in the topological property of the networks. Our application to transportation networks has been a kind of territorial benchmark for this novel approach, but the proposed method for detecting cores in communities through the optimization of the modularity function is quite general and can be applied to other networked systems.

Acknowledgments

The authors wish to thank The Atlanta Regional Commission for the availability of the regional travel demand model. AC acknowledges support from FET Open project 255987 FOC and CNR PNR project CRISIS Lab. Finally, GS wishes to thank Prof. Franco Meloni and the Dipartimento di Fisica of the University of Cagliari for the kind hospitality.

-
- [1] S. Fortunato, "Community detection in graphs," *Physics Reports*, **486**, 75–174 (2010).
 - [2] *Traffic Detector Handbook*, FHWA operations material, available at <http://www.fhwa.dot.gov/>.
 - [3] E.J. Mishan, *Cost-Benefit Analysis*, Praeger Publishers (1976).
 - [4] B.P. Pashigian, *Price Theory and Applications*, McGraw-Hill (1995).
 - [5] J. Sohn, "Evaluating the significance of highway network links under the flood damage: An accessibility approach," *Transportation Research A*, **40**, 419–506 (2006).
 - [6] R. Albert, H. Jeong and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, **406**, 378–382 (2000).
 - [7] R.V. Solé, M. Rosas-Casals, B. Corominas-Murtra and S. Valverde, "Robustness of the European power grids under intentional attack," *Phys. Rev. E*, **77**, 1–7 (2008).
 - [8] C.M. Schneider, A. Moreira, J.S. Andrade, S. Havlin, and H.J. Herrmann, "Mitigation of malicious attacks on networks," *Proc. Nat. Acad. Sci.*, **108**, 3838–3841 (2011).
 - [9] M.E.J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, **69**, 026113 (2004).
 - [10] S. Fortunato and M. Barthelemy, "Resolution limit in community detection," *Proc Natl Acad Sci, USA*, **104**, 36–41 (2007).
 - [11] M.E.J. Newman, "Modularity and community structure in networks," *Proc Natl Acad Sci, USA*, **103**, 8577–8582 (2006).
 - [12] V.D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, **10**, P10018 (2008).
 - [13] *Censimento generale della popolazione e delle abitazioni - matrice origine destinazione degli spostamenti pendolari della Sardegna*, Italian National Institute of Statistics (ISTAT) (1991).
 - [14] *The Travel Forecasting Model Set For the Atlanta Region - 2008 Documentation*, by the Atlanta Regional Commission (2008).
 - [15] A.G. Wilson, "Land-use/Transport Interaction Models," *Journal of Transport Economics & Policy*, **32**, 3–26 (1997).
 - [16] M. Batty, *Urban Modelling: Algorithms, Calibrations, Predictions*, Cambridge University Press (1976).